
# DATA WAREHOUSING & DATA MINING

# PROJECT REPORT

**SUBMITTED BY:**

| Student Name | Student ID |
|---|---|
| Mithun Chanda Shuvo | 19-39592-1 |
| Mehrab Hassan Pretum | 19-39719-1 |
| Syed Bishal Ahmed | 19-39584-1 |

**SECTION: B**

**DEPARTMENT: CSE**

**SUBMITTED TO:**

**COURSE TEACHER: TOHEDUL ISLAM**

**SUBMITTED DATE: 8/24/2022**

## Introduction:

Data mining is the method of extricating designs and other valuable data from expansive data sets. It's some of the time known as information disclosure in data or KDD. KDD is a significant process of identifying meaningful information and patterns in Data. The input is given to this process is data and output gives useful information from data. Some of the classification methods used in data mining include KNN, Naive Bayes, K-means clustering, Hierarchical Clustering and Decision Tree. I have chosen Dataset from Kaggle the dataset indicates that Stroke Prediction using different classifier. For Task1, we used supervised learning which is Naive Bayes and Knn algorithm. And the other one is Unsupervised learning which is K means clustering algorithm. By using that we can find the best suited classifier for the data set.

## Information about the data set :

Firstly, we have to understand that which is our Targeted feature and Others feature.

The Targeted feature is:

- ✓ Stroke

The other feature:

- ✓ Gender
- ✓ Age
- ✓ Hypertension
- ✓ Heart disease
- ✓ Ever married
- ✓ Work type
- ✓ Residence type
- ✓ Avg. glucose level
- ✓ Bmi
- ✓ Smoking status

There is total 4981 instances of these 11 attributes and all these instances were used for classification. Here are the graphical details of the attributes:

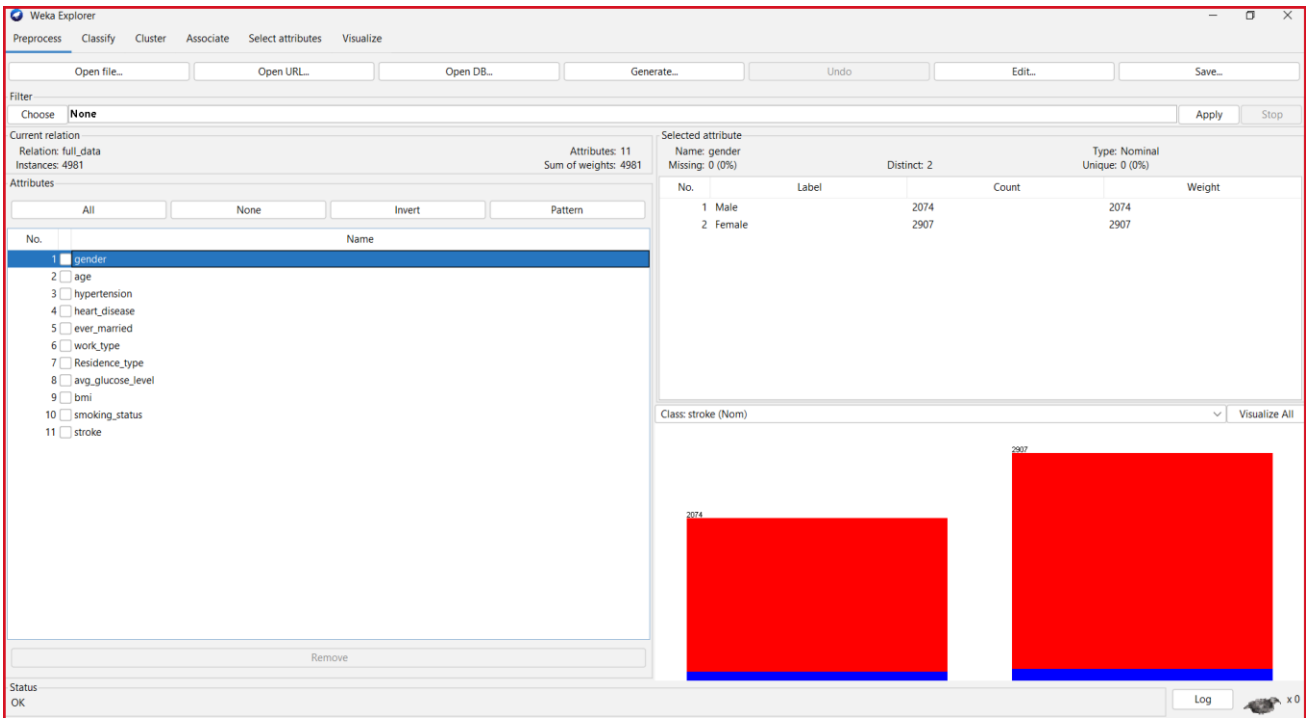



**Figure 1: Dataset Entry**

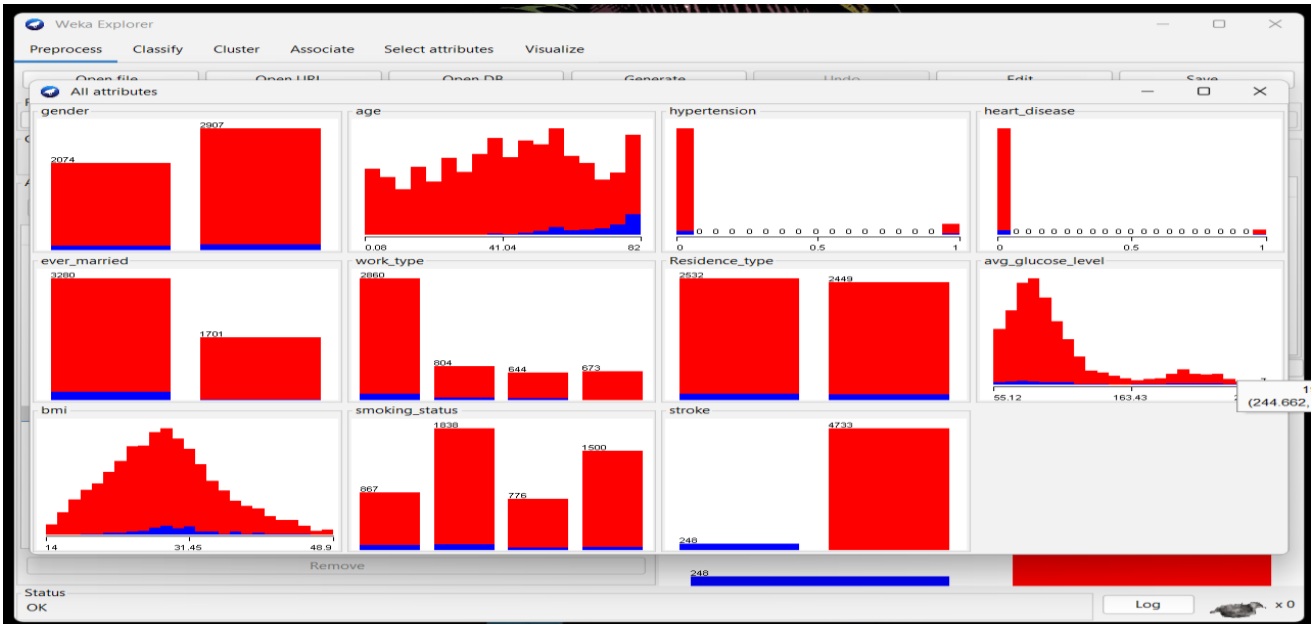



**Figure 2: All Attribute Visualization**

## For Supervised Learning:

**Applying Naive Bayes Classifier:** The Naive Bayes strategy could be a supervised learning algorithm for tending to classification issues that's based on the Bayes hypothesis. It is generally utilized in content classification assignments that require a huge training data set. In this data set we firstly used Naive Bayes classifier.

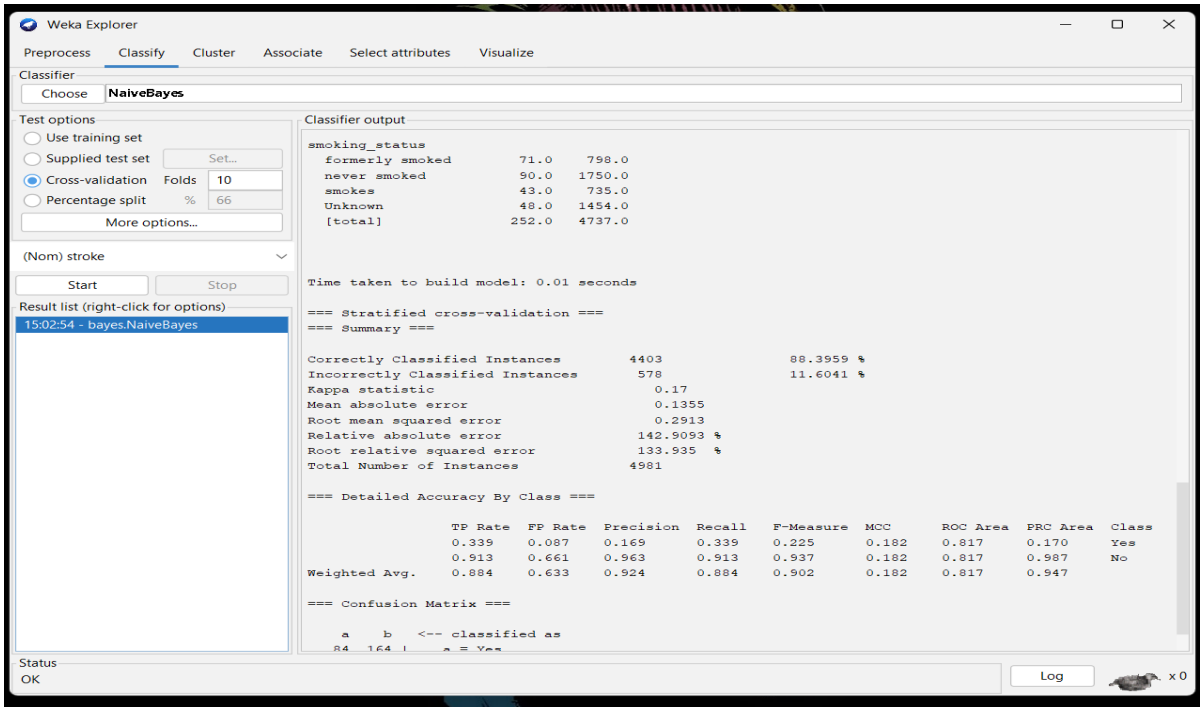



**Figure 3: Naive Bayes Classification**

**Applying KNN Classifier:** KNN classifier is used a supervised learning algorithm.it classifies data like new, unlabeled by analyzing the k number of nearest data points. The number of k depends on the data. Larger values k can reduce the noises. Accuracy of KNN can be affected by the noisy values or irrelevant attributes.

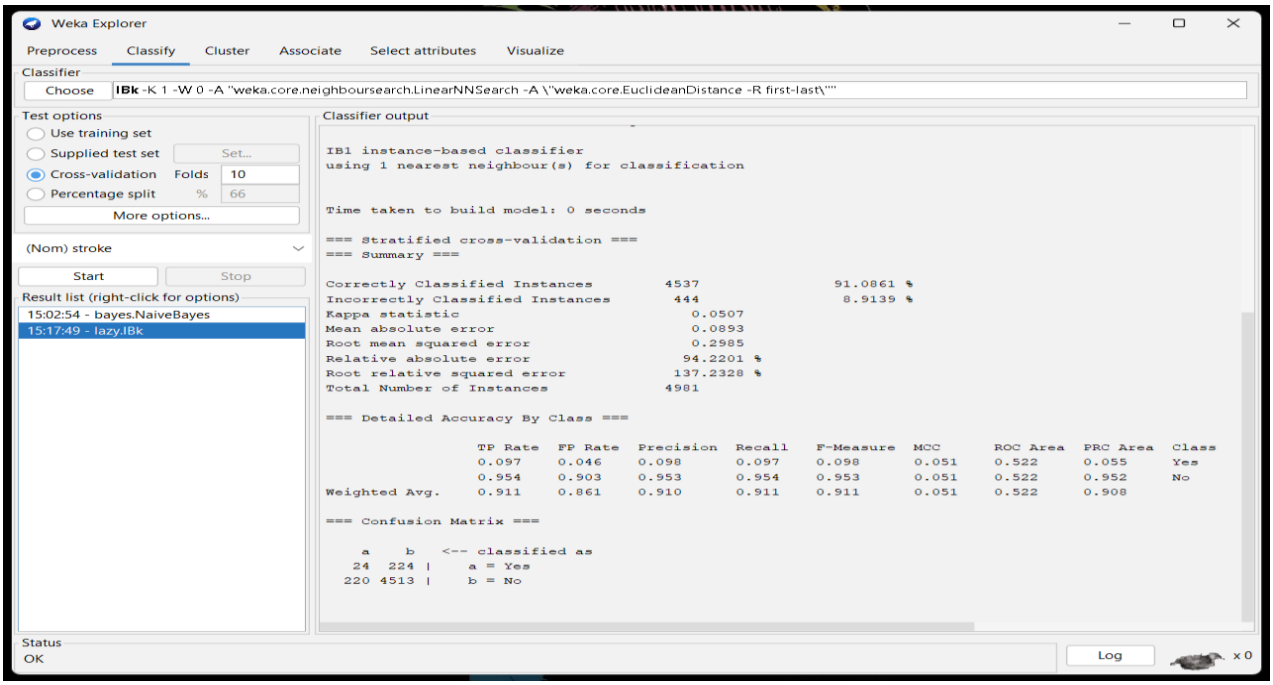



**Figure 3: KNN Classification**

After applying two types of classification, the highest percentage of correctly classified instances is for KNN classifier algorithm that is 91.08%. Another classifier is Naive Bayes that's correctly classified instance is 88.395%. Then Compare with this two classifier KNN classifier Value is greater than Naive Bayes Classifier. For this reason, KNN classifier algorithm is considered the best classifier for the data set.

**Choose KNN classifier algorithm:** The percentage of instances accurately classified is 91.08%. It would be a better classifier for the data set because it has the highest accurate value. One of the benefits of KNN classifier algorithm is that their outputs are simple to read and understand without the need for statistical expertise. KNN classifier algorithm demand less data preparation than other decision techniques. KNN classifier algorithm are more adaptable and simpler to use. This improves the accuracy of predictions.

## PREPARING TEST-DATA SET:

A test data set is a subset of training dataset. It is used to check the performance of the training dataset. The test dataset is used for assessment of the generalization error of the final dataset. After that we will get the idea about our algorithm's performance on the provided data.

If the test set contains $N$ instances of which $C$ are correctly classified, $C$ are correctly classified. Predictive accuracy, $P = C/N$. There are 30 instances in this prepared test data set.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | gender | age | hyperten | heart_di | ever_mar | work_typ | Residenc | avg_gluc | bmi | smoking_ | stroke |
| 2 | Male | 67 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly | Yes |
| 3 | Male | 80 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never sm | Yes |
| 4 | Female | 77 | 1 | 0 | Yes | Self-emp | Urban | 199.84 | 28 | formerly | Yes |
| 5 | Male | 78 | 0 | 0 | Yes | Self-emp | Urban | 218.46 | 26.8 | Unknowr | Yes |
| 6 | Female | 68 | 0 | 0 | Yes | Private | Rural | 211.06 | 39.3 | Unknowr | Yes |
| 7 | Female | 51 | 1 | 0 | Yes | Private | Urban | 88.2 | 28.4 | never sm | Yes |
| 8 | Male | 60 | 0 | 1 | Yes | Private | Urban | 91.92 | 35.9 | smokes | Yes |
| 9 | Male | 68 | 0 | 0 | Yes | Private | Rural | 233.94 | 42.4 | never sm | Yes |
| 10 | Female | 68 | 1 | 1 | Yes | Private | Urban | 247.51 | 40.5 | formerly | Yes |
| 11 | Male | 3 | 0 | 0 | No | children | Rural | 95.12 | 18 | Unknowr | No |
| 12 | Male | 58 | 1 | 0 | Yes | Private | Urban | 87.96 | 39.2 | never sm | No |
| 13 | Female | 8 | 0 | 0 | No | Private | Urban | 110.89 | 17.6 | Unknowr | No |
| 14 | Female | 70 | 0 | 0 | Yes | Private | Rural | 69.04 | 35.9 | formerly | No |
| 15 | Female | 52 | 0 | 0 | Yes | Private | Urban | 77.59 | 17.7 | formerly | No |
| 16 | Female | 32 | 0 | 0 | Yes | Private | Rural | 77.67 | 32.3 | smokes | No |
| 17 | Female | 79 | 0 | 0 | Yes | Govt_job | Urban | 77.08 | 35 | Unknowr | No |
| 18 | Female | 37 | 0 | 0 | Yes | Private | Rural | 162.96 | 39.4 | never sm | No |
| 19 | Female | 37 | 0 | 0 | Yes | Private | Rural | 73.5 | 26.1 | formerly | No |
| 20 | Female | 40 | 0 | 0 | Yes | Private | Rural | 95.04 | 42.4 | never sm | No |
| 21 | Male | 35 | 0 | 0 | No | Private | Rural | 85.37 | 33 | never sm | No |
| 22 | Female | 20 | 0 | 0 | No | Private | Urban | 84.62 | 19.7 | smokes | No |
| 23 | Female | 42 | 0 | 0 | Yes | Private | Rural | 82.67 | 22.5 | never sm | No |
| 24 | Female | 44 | 0 | 0 | Yes | Govt_job | Urban | 57.33 | 24.6 | smokes | No |
| 25 | Female | 65 | 1 | 0 | Yes | Private | Rural | 75.7 | 41.8 | Unknowr | No |
| 26 | Female | 49 | 0 | 0 | Yes | Private | Rural | 60.22 | 31.5 | smokes | No |
| 27 | Female | 59 | 0 | 0 | Yes | Private | Urban | 109.82 | 23.7 | never sm | No |
| 28 | Female | 25 | 0 | 0 | Yes | Private | Urban | 60.84 | 24.5 | never sm | No |
| 29 | Female | 67 | 0 | 0 | Yes | Govt_job | Rural | 94.61 | 28.4 | smokes | No |
| 30 | Female | 38 | 0 | 0 | No | Private | Rural | 97.49 | 26.9 | never sm | No |
| 31 | Female | 54 | 0 | 0 | Yes | Private | Rural | 206.72 | 26.7 | never sm | No |

**Procedure Of Test Dataset:** For the test data set we taken 30 instances. Here all the step shown.

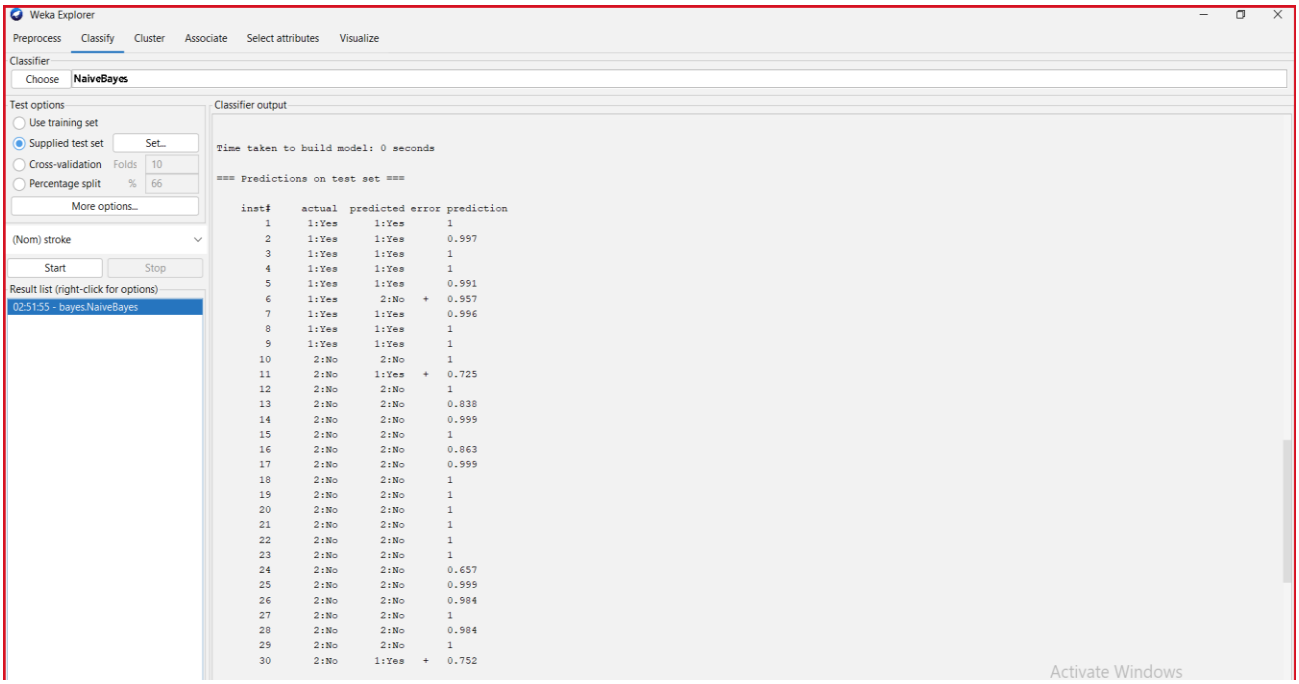



**FIGURE 4: Prediction Test Dataset**

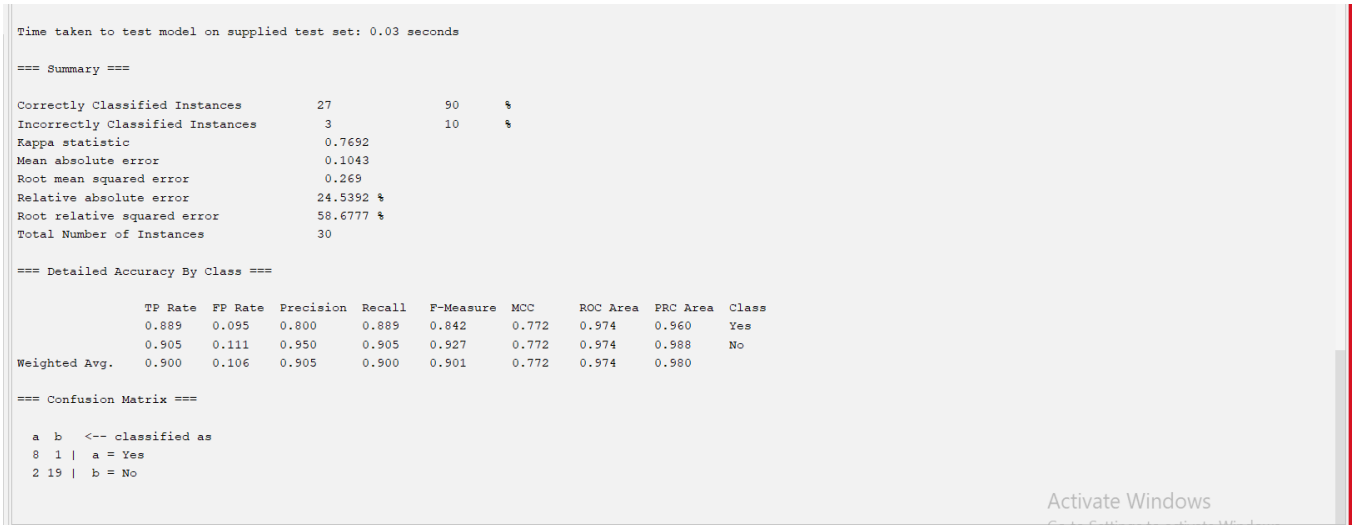



**Figure 5: Details of Test Dataset**

After completing Test Dataset, For 30 instances Correctly Classified Instances 27 and percentage is 90%. Incorrectly Classified Instances 3 and percentage 5%. Correctly classified Instance percentage are Good.

## For Unsupervised Learning:

**K-Means Clustering Algorithm:** K-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities.

In this portion we take another dataset that is unsupervised. Here total 13 attribute and 178 instance.

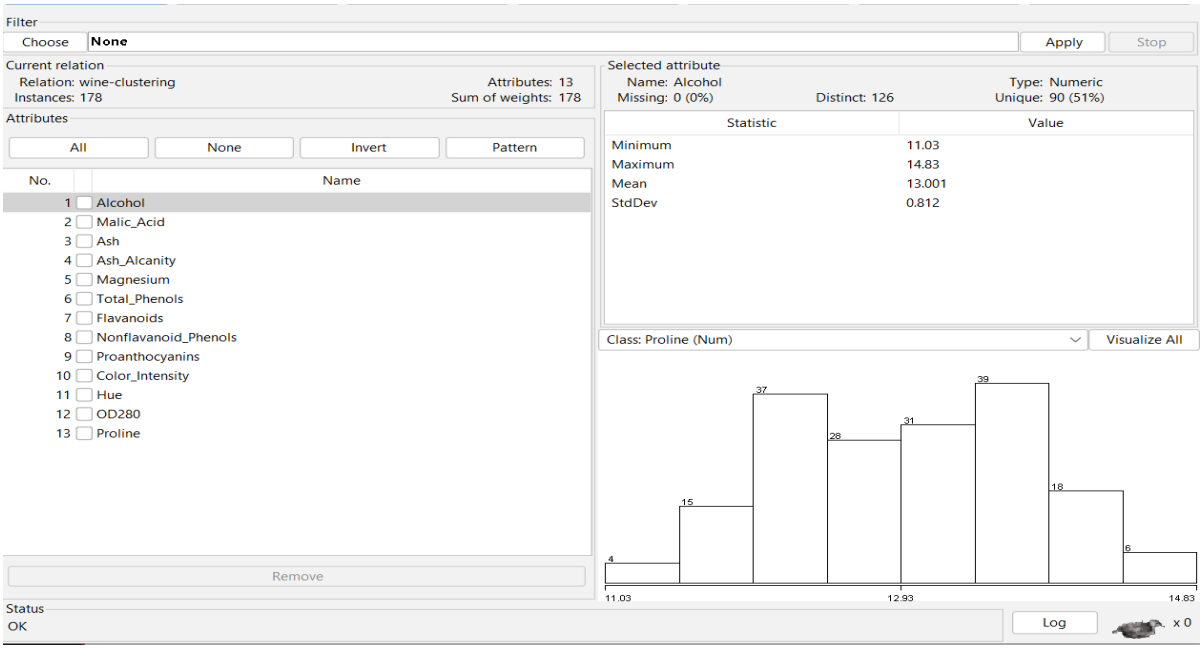
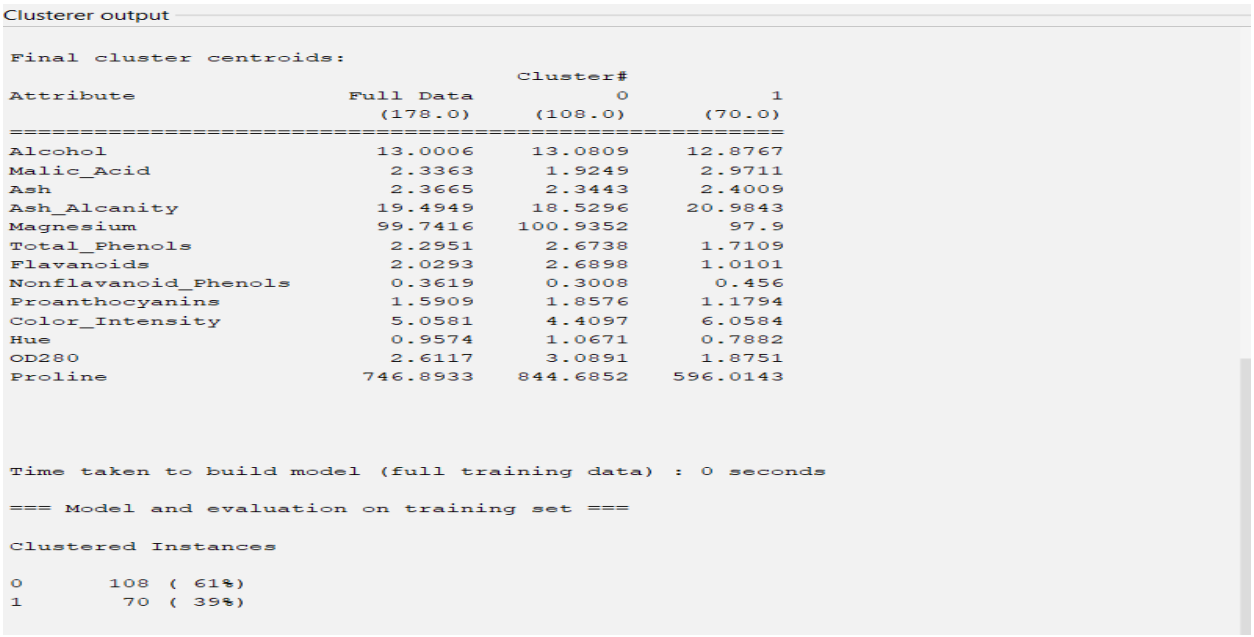



**Figure 6: Unsupervised Dataset Entry**



Clusterer output

```
Final cluster centroids:
                                Cluster#
Attribute               Full Data           0           1
                          (178.0)     (108.0)      (70.0)
============================================================
Alcohol                   13.0006     13.0809     12.8767
Malic_Acid                 2.3363      1.9249      2.9711
Ash                        2.3665      2.3443      2.4009
Ash_Alcanity              19.4949     18.5296     20.9843
Magnesium                 99.7416    100.9352        97.9
Total_Phenols              2.2951      2.6738      1.7109
Flavanoids                 2.0293      2.6898      1.0101
Nonflavanoid_Phenols       0.3619      0.3008       0.456
Proanthocyanins            1.5909      1.8576      1.1794
Color_Intensity            5.0581      4.4097      6.0584
Hue                        0.9574      1.0671      0.7882
OD280                      2.6117      3.0891      1.8751
Proline                  746.8933    844.6852    596.0143



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       108  ( 61%)
1        70  ( 39%)
```

**Figure 7:Details of Unsupervised Dataset**

**Discussion:**

The purpose of this report was to find a suitable classifier for stroke prediction which will classify the stroke as accurately as possible and will be able to predict the class from the data set of stroke prediction. After applying two different classifier which are Naive bays and KNN classifier algorithm which we used for supervised learning. And the other clustering method we used is k means clustering which we used for unsupervised learning. From the supervised learning our best-chosen classifier for the data set is KNN classifier algorithm with 91.0861 % accuracy. We prepared test set with 30 instances was used to test the model and finally the model accuracy is 90% for the prepared test data set. Creating training and test data set is an important concept in data science as it is used to improve generalization and minimizing over fitting. Then for the unsupervised learning we use k means clustering algorithm. As we see in the picture for cluster 0, the instances are 108 and the percentage is 61% and then for the cluster 1, the instances are 3277 and the percentage is 39%.

**References:**


- **https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset**
- **https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering**